CLEAN VERSION

INTERNET CACHING SYSTEM

Technical field of invention

The present invention relates to a method, a system and a server for caching Internet information content.

Background to the invention and Prior Art

Over the last few years, Internet has developed into the fastest growing means of communication and is predicted to be the main system for distributing information in the future. The Internet has become increasingly popular and the number of users is growing at a magnificent rate. Anybody can provide any information - text, pictures, audio and video - on the net where it can be retrieved by users anywhere in the world. This makes for the incredible success of the Internet and its currently most used feature, the World Wide Web, WWW.

However, the popularity of the Internet, and especially the use of the Internet application World Wide Web, is putting the net under enormous capacity pressure. Traffic on the Internet doubles about every three months, with no signs of slowing down. This kind of exponential growth will eventually slow down, but several factors indicate that it is still far from any slowdown in growth.

Today, only a small fraction of all PCs have access to the Internet. It is believed that most of those who have made and will make the investment in a PC will eventually want access to the Internet. Furthermore, communication speeds between the user and the Internet infrastructure are increasing rapidly. Technologies such as ISDN, Cable TV modems and xDSL are emerging. Moreover, manufacturers of household electronics are launching low cost devices to give Internet World Wide Web access through television sets.

These and other factors are all contributing in making it increasingly difficult to access information on the Internet, since the system is basically being overloaded. In many parts of the world, the lack of sufficient bandwidth is a major problem. Also, passing information between USA and Europe is limited by the transmission capacity of the transatlantic link. The transmission capacity is simply too small to cope with the exploding Internet use.

Basically, there exists two conventional solutions to this problem. A first solution is to add more bandwidth and switching capacity, which is being done, but which carries huge costs and also faces significant technical problems.

A second solution is to use so called caching techniques. Caching basically means to keep track of the Internet traffic and to keep copies of the most frequently accessed World Wide Web files at a location closer to the user than the original files. Hence, this involves the storing of a local copy of a World Wide Web Site of interest. For example, a copy of CNN's USA-based homepage may be temporarily stored in a cache at a European "proxy" location, whereby European Internet users may access the CNN homepage without having to use the transatlantic communication link, thereby both attaining quicker access and at the same time lowering the load on the Internet. The total World Wide Web information content itself is getting to be unmeasurable, but only a relatively small subset of all this information accounts for a huge proportion of what is actually viewed. Depending on the size and homogeneity of a user community, about 10-20 Gigabytes of cache storage will (spring 1997) reduce the traffic by 30-50% in that community.

However, in spite of current caching techniques, in many parts of the world, such as South and East Europe, South America, India and East Asia, the bandwidth capacity of the available communication systems are far too low to provide a desired usability of the Internet.

Summary of the invention

An object of the invention is to reduce the capacity constraints on the Internet by providing a caching solution which reduces access time on the Internet and which allows growth to a significantly larger amount of users and information traffic on the Internet, especially in areas where the bandwidth capacity is low.

According to the present invention, the objects are achieved by a method, a system and a server for caching Internet information content according to the appended claims.

The invention is based upon the insight that a linguistically and culturally homogenous geographical area often differ in size and shape from the economically and technically optimal region to be served by a single cache server. Therefore, such an area is served by a set of geographically distributed servers. However, the invention is further based upon the insight that, under the assumption that users in such a linguistically and culturally homogenous geographical area will have homogenous preferences regarding visited information sites, information cached in one such server due to a user request should also be cached in the other servers within said area, the basic principle being that if one user is interested in a certain piece of information, it is likely that so will other persons within the same linguistically and culturally homogenous geographical area. The larger the end user community, the larger the probability that someone else within the community has requested any given file.

Hence, according to the invention, a set of geographically distributed cache servers serving different geographical regions but having a common relation, such as the relation of serving a culturally and/or linguistically defined user group or area, is updated with essentially the same information whenever one of said cache servers retrieves information due to an information request from an end user or the like.

Such cultural and linguistical homogeneity may be defined by the boundaries of a nation, the boundaries of a specific language, the boundaries of a specific religion, the boundaries of a certain level of technology or development, the boundaries of a certain economical area, as well as combinations thereof, or the like.

Hence, according to an embodiment of the invention, the set of geographically distributed cache servers are distributed within a linguistically and culturally defined global geographical area or a geographically defined community. Furthermore, each of the cache servers is preferably arranged to serve a respective local region within said geographical area or community.

For example, the German speaking parts of Germany, Austria, Switzerland and Italy may be seen as defining a cultural and linguistically homogenous global area. Therefore a set of cache servers operating according to the invention is distributed within this area. Among this set of servers, provided within the global area defined above, one (or more) may, for example, be arranged to serve a first region covering the northern part of Germany, a second one (or more) to serve a second region covering the eastern part of Germany, a third one (or more) to serve a third region covering the central part of Germany, a fourth one (or more) to serve a fourth region covering Switzerland and the northern part of Italy, and finally a fifth one (or more) to serve a fifth region covering Austria.

Note that a defined geographical area according to the invention does not necessarily imply a homogenous, continuos area encompassed by a closed boundary, but may just as well mean a set of geographically separated sub-areas together defining an area according to the invention.

According to estimations made by the inventor, in some parts of the world, such as India and South America, the bandwidth capacity of the telecommunication network is so small, that it is actually more efficient to distribute a specific package of information to all cache servers within the defined global area based upon just one information request to the Internet from one single user, than

to have another user search the Internet for the same package of information a second time. However, rules for determining when to perform distributed caching according to the invention or not, may be set in different ways depending on the actual application environment, as will be discussed below.

The distribution of information to the set of cache servers according to the invention may be performed in many different ways. According to one aspect, the Internet could be used. However this would put even more strain on the transfer capacity of the net. Therefore, according to a preferred embodiment of the invention, the information is distributed to said set of cache servers using multicast communication, preferably using dedicated data channels separated from the Internet communication itself.

According to a preferred embodiment, said information is distributed to said set of cache servers using a satellite link. Each of the servers then preferably have access to an uplink for sending said information, and all servers receive a downlink which provides information sent from any one of the servers. The hardware and software for handling such distribution may be provided in association to the cache server itself, or may be provided by for example an Internet Service Provider or a Satellite Communication Link Provider. Similarly, the software for controlling the information flow and for performing decisions within the system may be provided at the cache servers, at the Internet Service Providers or wherever convenient, as will be clear to a person skilled in the art. If a satellite system is used, the satellite link may also be used for transferring or relaying requests to other parts of the globe, for example from Europe to USA and vice versa.

According to an alternative embodiment, all cache servers within the system are connected via data channels to a central control unit. Such a central control unit may for example control the decisions as to whether or not information should be cached and as to how it is to be distributed and so on.

Depending on the size of the community, the capacity of the cache servers, the available communication capacity of the Internet, the objects of the cache provider, and so on, the distributed caching technique according to the invention may be arranged to distribute, or not distribute, information based on different rules. According to a preferred embodiment of the invention, only requests or frames referring to a specific communication format or application, corresponding to the type of information service provided on the local storing means, is cached. For example, only so called TCP-queries to the WWW port. According to another example, the cache server provider may decide only to cache from specific addresses (e.g. only .com or .org addresses), only after a certain number of requests have been made within the system (e.g. after three separate requests), or the like.

The processing means necessary for performing such decisions may be arranged at each cache server, at a central control unit as mentioned above, at an Internet Service Provider or similar location working in association with the cache servers, or the like.

Also request relating to an Information Content Provider located in a region served by one of the cache servers in said set of cache servers may be handled in different ways according to different embodiments of the invention, as will be discussed in the following.

According to a first embodiment, it is assumed that a local information request from an end user to an Internet content provider, both being located within the same region, only is of local concern. Therefore the information shall: a) not be cached at all, since a local user will always have easy access to a local Internet Content Provider; or b) only be cached in the local cache server and not be distributed to the entire set of geographically distributed cache servers, since the request probably only is of local concern. Hence, in this context, the term "local" refers to features residing within one of said regions.

According to a second embodiment, it is assumed that an information request from an end user to an Internet content provider located within one of said regions shall not be cached in the cache server serving that region, but shall instead only be distributed to all other cache servers within the system, since a local user will always have easy access to a local Internet Content Provider (hence no need for local caching) but a user outside the region will not have the same easy access to the local Internet Content Provider (hence the need for caching outside the region).

According to another embodiment, it is assumed that the cache server capacity is large enough to handle all requests, therefore there is no need for handling information relating to local Internet Content Providers different than other information and, consequently, the caching of information in one cache server within the system shall result in the caching of the same, information in all cache servers within the system.

According to yet another embodiment of the invention, depending on the choice of rules for decisions within the system, there is provided a list of Internet Content Provider addresses to be excluded from caching. Hence, each cache is arranged to check incoming information requests against said list to see whether or not the request shall lead to caching.

Depending on the desired utility, the cache servers will be arranged to store information for a given amount of time. For example, the caches may operate on the principle of first-in-first-out. According to another example, certain kind of information may be provided with update requirements, which will define the longest time interval that the copied information may be used to ensure that the information is still correct. Of course, regarding this and other aspects, the invention may be combined with different types of conventional caching techniques, as is understood by one skilled in the art.

Furthermore, each of said cache servers may be constituted by several sub-servers connected, preferably via high speed switching means, to a cache manager controlling the operation thereof. Further, there is no limit to the number of cache servers that may be used. Also, each server may be used by more than one Internet Service Provider.

In the event of a failure of one of said cache servers, it can rapidly be bypassed by the system to restore some service while the cache server is being repaired or replaced. The cache functionality may then be provided from another server in another region within the system. An advantage of the invention is that the auxiliary cache server then provides the same stored content as the bypassed serves, thus making sure that the cached information is still relevant to the end users in the region of the bypassed server.

Although the description of the invention has been made with respect to the Internet communication system, it is understood by those skilled in the art that the invention also may be used in other similar types of global information communication systems which may be found in the future and which may show similar kinds of communication problems. The invention is therefore of course not limited to the Internet application.

Also, even though the description further has been made mainly with respect to the World Wide Web application, it is to be understood that the invention may as well be used for many other kinds of net applications and is not limited thereto.

Brief description of the drawings

Further aspects, features and advantages of the invention will become clear from the following description of preferred exemplifying embodiments with reference to the accompanying drawings, in which:

Fig 1 schematically shows an embodiment of an internet caching system according to the present invention;

Fig 2 schematically shows a cache server arrangement forming a part of the system shown in Fig 1;

Fig 3 schematically shows an alternative cache server arrangement;

Fig 4 schematically shows yet another alternative cache server arrangement;

Fig 5 schematically shows a flow chart of the operation performed by the cache server in Fig 4;

Fig 6 schematically shows a flow chart of the operation performed by the interceptor in Fig 3;

Fig 7 schematically shows another embodiment of an internet caching system according to the present invention; and

Fig 8 schematically shows yet another embodiment of an internet caching system according to the present invention.

Detailed description of preferred embodiments

A first embodiment of an internet caching system according to the invention will now be described with reference to Fig 1. In Fig 1, a preferably linguistically and/or culturally defined geographical area 200 is shown encompassing four geographical regions A, B, C and D. For example, the area 200 could be the British islands, and the regions A, B, C and D could then be Ireland; Wales; Scotland and Northern England; and Southern England, respectively.

In each of the regions A, B, C, and D, there is at least one Internet Service Provider 110a, 110b, 110c, and 110d, respectively, providing the respective region with access to the Internet 100. Each service provider 110a, 110b, 110c, 110d, hence provides Internet connection to a

number of end users 120a, 120b, 120c, and 120d, respectively, and a number of Internet content providers 130a, 130b, 130c, and 130d, respectively. It is understood that, even though the end users 120 and content providers 130 are shown in Fig 1 as being separate features, an end user may be a content provider as well.

Each service provider is connected to at least one Internet cache server 140a, 140b, 140c, 140d, respectively. Each cache server 140 hence provides cache functionality to the end users 120 within the respective region. Also, each cache server 140a, 140b, 140c, 140d, has a connection to a satellite disc 150a, 150b, 150c, and 150d, respectively, and may broadcast information to all other cache servers within the area 200 using a satellite 160.

An example of the operation of the system shown in Fig 1 will now be described. When, for example, a service provider 110a in region A receives an information request from an end user 120 relating to information provided by a content provider 130e located in a region E, which in this case lies outside the area 200 (for example, the region E is North America), the service provider 110a checks its associated cache server 140a to see if an updated copy if the requested information is stored therein. If so, the requested information is derived from the cache server 140a and is then delivered to the end user. If, however, no copy of the requested information is stored in the cache server 140a, the service provider 110a relays the request to the original content provider 130e via the Internet 100. When the requested information is returned from the content provider 130e via the Internet 100, the service provider 110a delivers the information to the end user and updates the cache server 140a with the received information. As a part of the updating operation, the updated information is sent via the satellite disc 150e and the satellite 160 to the cache servers 140b, 140c, 140d in the other regions B, C, and D, thereby updating all servers with the same information. After this updating has been made, no

user 120 within the area 200 needs to access the original content provider 130e via the Internet, but can instead use the copy stored in the respective cache server. If any cache server is already full when updating, the information having been stored the longest period of time (since last accessed by an end user) will be deleted from said server. Also each server will use on of several methods to ensure that the information stored in each cache server are accurate and up to date, hence providing reliable copies of the original sites.

In the following figures, for ease of description, elements having similar functions as those described with reference to Fig 1 will be designated with the same numerals.

A cache server arrangement forming part of a system of the kind shown in Fig 1 will now be described with reference to Fig 2. In Fig 2, three end user terminals or stations 120 are connected to a multiplexing unit 170, which may be a modem pool or a LAN router depending on the type of user system. The multiplexing unit 170 is connected to an interceptor 110 arranged at an Internet Service Provider providing access to the Internet 100, The multiplexing unit 170 multiplexes the communications to and from the user terminals 120.

According to this arrangement, a user 120 wishing to access information provided for example as a World Wide Web homepage or site by an Internet content provider (not shown), located at a original site of publication somewhere on the Internet 100, generally sends an information request, specifically a HTTP request to the WWW-port, via the multiplexing unit 170 to the Internet service provider 110, where the request is routed out on the Internet 100 to finally reach the content provider. The content provider then answers the request, sending the requested information via the Internet 100 back to the user 120 via the service provider 110.

However, according to the arrangement shown in Fig 2, the information request from a user 120 is examined by the interceptor 110 at the service provider. The interceptor then decides

whether or not the requested information, for example the information provided by the content provider, exists in a copy or similar form on a local cache server 140 arranged in direct communication with the interceptor 110. If a copy or the like of the requested information is stored at the local server 140, the request from the user is re-routed to the server 140, which will then return the requested information to the user. In this case, the information will already exist on all cache servers within the area 200 shown in Fig 1, and, hence, there is no need for distributing the information to other caches.

However, if the interceptor 110 determines that no copy or the like of the information of interest exists on the local server 140, then the information request from the user 120 is sent on to the Internet 100 as in the conventional form of Internet communication. The interceptor will then monitor the response from the Internet 100 to the user 120 and update the cache server 140 with the requested information as it is received.

The caching of new or updated information in the cache server 140 is, according to the invention, accompanied by the distribution of the new or updated information to other cache servers within the area 200 shown in Fig 1. This distribution is enabled by either the cache server 140 or the interceptor 110 at the service provider, for example, and uses a satellite uplink to the satellite 160 to reach other cache servers within the defined area. Similarly, when receiving information relating to new or updated information from other cache servers within the area, the cache server 140 will receive such information via a downlink from the satellite.

Note that the interceptor 110 in this example only intercepts communication packages of certain specific types. For example, only World Wide Web information requests from the users 120 are intercepted. Other types of communication services, such as e-mail or the like, are left undisturbed by the interceptor 110.

An alternative cache server arrangement, similar to the one shown in Fig 2, will now be described with reference to Fig 3. The arrangement in Fig 2 differs from the one shown in Fig 2 in that the interceptor 110 at the service provider is connected to not only the cache server 140, but also to a quick access cache memory 112. In comparison to the cache server 140 according to the invention, for example providing several hundred Gigabytes of storage, the quick access cache memory will in this case provide a much smaller storage capacity, for example around one Gigabyte of storage. The cache memory 112 is therefore used to store essentially all information passing the interceptor 110, but consequently, each package of information will only be stored a comparatively small amount of time.

In Fig 3, the quick access cache memory, comprising a memory area 112 and a look-up table 111 showing which information is currently stored in said memory area, is used as a working memory of the interceptor 110. Hence, when receiving an information request from a user 120, the interceptor will first check its quick access cache memory 111, 112 for the requested information. The cache server 140 will then be accessed only if the requested information cannot be found on the cache memory 112. The operation follows the operation described with reference to Fig 2. However, in Fig 3, all information relating to every information request from the users 120 will be temporarily stored in the quick access cache memory 112, either information received from the Internet 100 or information received from the cache server 140, whereas the cache server will only be updated when new information is received from the Internet 100. Furthermore, the cache memory may be used to cache information of other types than the information being stored in the cache server 140. For example, in addition to storing WWW information, as the cache server 140 does, the cache memory may also store information relating to other types of queries. Also, information stored in the cache memory 112 will not automatically be distributed to other caches within the area 200 in Fig 1.

Another alternative cache server arrangement, also similar to the one shown in Fig 2, will now be described with reference to Fig 4. In Fig 4, all WWW requests will be relayed form the service provider 110 to the cache server 140. The cache server will then answer the request, if the requested information is stored therein, or handle the information retrieval from the internet 100, either using a satellite link or some other connection 141 to the Internet. When receiving the requested information, the cache server will 140 update itself, as well as other caches via the satellite 160, according to its own decisions, and deliver the information to the end user 120 via the service provider 110.

Note in Fig 4 that the service provider may still handle other types of services, such as email or the like, on its own, without relaying such information packages via the cache server.

An example of the operations performed by the cache server 140 in Fig 4 will now be described with reference to the flow chart in Fig 5. After activating the operation in step S1, the cache server receives, in step S2, an information request from an end user relating to a specific content provider. In step S3, the cache server then determines whether or not a copy of the requested information is stored therein. If so, the requested information is delivered to the end user in step S4 and the operation is ended in step S10. (Of course, the end step S10 may cause a return to the starting step S1).

However, if it is determined in step S3 that no copy of the requested information is stored in the cache server, the process continues to step S5, wherein the cache server sends a request for said information to the designated content provider via the Internet. The cache server then receives the requested information in step S6 and delivers it to the end user in step S7. In step S8, the received information is stored in the cache server, and then, in step S9, the information is distributed to all other cache servers within the area 200 shown in Fig 1. Hence, the received new, previously not stored

information will then be stored in essentially all cache servers within the area and thus be easily accessed by all users within the area.

An example of the operations performed by the interceptor or service provider 110 in Fig 3 will now be described with reference to the flow chart in Fig 6. After activating the operation in step S11, the interceptor 110 reads, in step S12, an information request from an end user directed to a specific content provider. In step S13, the interceptor 110 checks the content table 111 of the quick access cache memory 112 to see if a copy of the requested information exists therein. If so, the operation continues to step S15 described below. If not, the operation continues to step S14, in which the interceptor sends a query to the cache server 140 to see if a copy of the requested information exists therein. If the answer is yes, the process continues to step S15, in which the request from the end user is intercepted, and to step S16, wherein the requested information is delivered, either using the cache memory 112 or the cache server 140. However, if the answer from the cache server is no, the process continues to step S17, wherein the interceptor determines not to intercept the request from the user, which is instead passed on to the designated content provider via the Internet. The process then continues to step S18, in which the interceptor monitors the answer sent from the content provider to the end user. When such information passes the interceptor on its way to the end user, the interceptor, in step S19, stores the previously unattainable information in the cache server 140 and distributes the information to all other cache servers within the area 200 in Fig 1. Hence, the received new, previously not stored information will then be stored in essentially all cache servers within the area and thus be easily accessed by all users within the area. Note that every time information is received by the cache server from the Internet, a copy is preferably temporarily stored in the quick access cache server 112.

Another exemplifying embodiment of an Internet caching system according to the present invention, similar to the one shown in Fig 1, will now be described with reference to Fig 7. In Fig 7, the

system differs from the one shown in Fig 1 in the answer from the content provider 130e, in this case being located in the region outside the system area 200, relating to an information request originally sent by an end user 120a, will not be transmitted back to the user via the Internet 100. Instead, the answer is transmitted directly via the satellite 160 to the servers 140a, 140b, 140c, and 140e, thereby automatically updating all cache servers 140a. The cache server 140a and the service provider 110a then handle the delivery of said information to the end user 120a.

Yet another exemplifying embodiment of an Internet caching system according to the present invention, similar to the one shown in Fig 1, will now be described with reference to Fig 8. In Fig 8, the cache servers within the area 200, each serving a respective region A, B, C, and D, is connected via data channels to acentral control unit 190. The central control unit 190 controls the decisions as to whether or not information shall be cached in the respective cache servers and as to how it is to be distributed to other caches within the area 200. In this case, information may be distributed between the different caches either using the satellite link or using the data channel connection to and via the central control unit 190.

Furthermore, with reference primarily to Fig 1, 7 and 8 above, it is understood that decisions regarding how to handle the caching of information relating to "local" requests, i.e. for example an information request from an end user 120a relating to a content provider 130a located within the same region A, which has been discussed above in the summary of the invention, the different schemes of operation, including the needed decision-making, may be implemented either at each of the cache servers 140, at each of the service providers 110, or at a central control unit 190, for example.

It is understood that the construction and function of the elements described with reference to the drawings will become apparent for those skilled in the art.

Even though the invention has been described with reference to specific exemplifying embodiments, many different alterations, modifications and the like will become apparent to those skilled in the art. The described embodiments are therefore not intended to limit the scope of the invention, as defined by the appended claims.

.HrtLib1:417707.1 01/17/03